Theoretical Computer Science 1 (1975) 95–101. © North-Holland Publishing Company

A COMBINATORIAL PROPERTY OF CODES HAVING FINITE SYNCHRONIZATION DELAY

Antonio RESTIVO

Laboratorio di Cibernetica del CNR, Arco Felice, Napoli, Italy

Communicated by M. Nivat
Received 15 November 1974

Abstract. Necessary and sufficient conditions are given under which a code has finite synchronization delay. The proof of the theorem makes use of three lemmas involving combinatorial properties of the free monoid related with the synchronization delay. A corollary of these lemmas (Property 3) gives also an upper bound on the synchronization delay of a code A, which linearly depends on the number of states of the minimal automaton recognizing A^*.

1. Introduction

Let X be a finite set (alphabet) and X^* the free monoid generated by X. A subset A of X^* is a *variable length code*, or simply a *code*, iff the submonoid A^* generated by A is free with basis A, or, alternatively, iff, for any one-to-one mapping ϕ of a set Y onto A, ϕ can be extended to a one-to-one homomorphism ϕ^* of the free monoid Y^* into X^*.

Following Schützenberger [12] we give the following definition.

Definition 1. Let A be a code. A pair (a, a') of words in A^* is *synchronizing* iff for all $f, f' \in X^*$, the relation $faa'f' \in A^*$ implies $fa \in A^*$ and $a'f' \in A^*$. A code A has *finite synchronization delay* iff there exists a natural number q such that every pair of words in A^q is synchronizing. The least q for which this condition is satisfied is called the *synchronization delay* of A.

Intuitively, (a, a') is a synchronizing pair if the word aa' determines without ambiguity a factorization of the word $faa'f'$ in elements of A, regardless of the "context" (f, f') in which aa' appears. A code A has then finite synchronization delay if a message, written in the code A, can be uniquely decoded (that is factorized into words of A) by examining only a finite number of elements of A, from any starting point.

The codes of Definition 1 can be considered the natural extension of other families of codes known in the literature, as the "locally parsable codes" [6] (in which A

is finite) and the "comma free codes" [4] (in which all the words of A have the same length). Some enumeration problems on these codes are solved in [3].

It is noteworthy that the notion of code having finite synchronization delay plays a roe in some problems of automata theory and in particular in the algebraic characterization of subfamilies of the rational languages [14, 9].

In this paper, extending a result obtained in [9], we relate the codes having finite synchronization delay with another remarkable family of codes (Definition 2) which appears in the factorizations of free monoids [13] and can be utilized in the construction of the basis of free Lie algebras [11, 15, 16].

Definition 2. A code A is *very pure* iff for all $u, v \in X^*$, the relations $uv \in A^*$ and $vu \in A^*$ imply $u \in A^*$ and $v \in A^*$.

Remark. The name "very pure" has been first used in [8] to denote a subfamily of the *pure* codes, introduced in the same paper. (A submonoid P of X^* is *pure* iff for all $f \in XX^*$ and for any natural number n, $f^n \in P$ implies $f \in P$. A code A is pure iff A^* is a pure submonoid.) The notion of a pure submonoid is borrowed from the theory of groups, where a subgroup which satisfies the above condition is called pure.

Definition 2 can be interpreted as follows. Recall that two words f, g of a free monoid X^* are *conjugate* [5] iff there exist $u, v \in X^*$ such that $f = uv$ and $g = vu$; if neither u nor v is the empty word, then f and g are *strictly conjugate*. Now let $A \subset X^*$ be a code and let ϕ be a one-to-one mapping of an alphabet Y onto A, extended to a one-to-one homomorphism ϕ^* of the free monoid Y^* into X^*; consider the inverse isomorphism $\psi = (\phi^*)^{-1}$ of A^* onto Y^*. Then A is very pure iff the hypothesis that f and g are strictly conjugate in X^* implies that $\psi(f)$ and $\psi(g)$ are also strictly conjugate in Y^*.

By the above interpretation, the following properties are easily verified.

Property 1. Any subset of a very pure code is itself a very pure code.

Property 2. Let $A \subset X^*$ be a very pure code and let η be a one-to-one homomorphism of X^* into a free monoid Z^* on a set Z, such that $B = \eta(X) \subset Z^*$ is also a very pure code. Then the compound code $A \otimes B = \eta(A) \subset Z^*$ (see [7]) is very pure.

Before stating the main result, let us recall some fundamental concepts from automata theory [2]. If L is a subset of X^*, the right equivalence ρ_L, induced by L, is defined as follows on the words of X^*:

$$f \rho_L g \text{ iff for all } h \in X^* \{fh \in L \Leftrightarrow gh \in L\}.$$

A subset L of a free monoid is *rational* (or regular) iff the right equivalence ρ_L has

finite index; this index gives also the number of states of the minimal automaton which recognizes L.

The main result of the paper is the following:

Theorem. (a) *Every code with finite synchronization delay is very pure.*

(b) *Every very pure code A, that is rational and that satisfies the condition $F(p)$: $A \cap X^*A^pX^* = \emptyset$, for some natural number p, has finite synchronization delay.*

If we suppose the code to be *finite*, the restrictive hypotheses of (b) are trivially verified and we obtain one of the results of [9]. Note that those restrictive hypotheses are indeed necessary, as shown by the following examples.

Example 1. Consider $X = \{x, y, z\}$ and $A = \{x\} \cup (yx^*z)$. A is very pure and rational, but, for any natural number q, the pair (x^q, x^q) of elements of A^q is not synchronizing. Indeed $yx^{2q}z \in A^*$ and $yx^q \notin A^*$.

Example 2. Consider $X = \{x, y\}$ and $A = \{(x^{2n-1}y^{2n}) \cup (y^{2n}x^{2n+1}) | n \in \mathbb{N}\}$. A is very pure and satisfies the condition $F(p)$, but, for any natural number q, the pair

$$(y^2x^3y^4x^5 \dots y^{2q}x^{2q+1}, y^{2q+2}x^{2q+3}y^{2q+4}x^{2q+5} \dots y^{4q}x^{4q+1})$$

of elements of A^q is not synchronizing. Indeed

$$xy^2x^3y^4 \dots y^{2q}x^{2q+1}y^{2q+2}x^{2q+3} \dots y^{4q}x^{4q+1}y^{4q+2} \in A^*,$$

but

$$xy^2x^3y^4 \dots y^{2q}x^{2q+1} \notin A^*.$$

In the course of the verification of the theorem, we shall also obtain the following:

Property 3. Let A be a rational code that satisfies the condition $F(p)$ for some p. If A has synchronization delay q (finite), then $q < 2p(s+1)$, where s is the number of states of the minimal automaton which recognizes A^*.

2. Proof of the theorem

We first recall a basic result on free submonoids and some of its elementary consequences.

Proposition (Schützenberger [10], see also [1]). *A necessary and sufficient condition for A^* to be free and of basis A is that for all $f \in X^*$, $fA^* \cap A^* \neq \emptyset$ and $A^*f \cap A^* \neq \emptyset$ implies $f \in A^*$.*

Corollary 1. *Let A be a code. For all $u, v \in X^*$, $uv \in A^*$ and $vu \in A^*$ implies $u \in A^*$ and $v \in A^*$ or $u \notin A^*$ and $v \notin A^*$.*

Proof. If $uv, vu, u \in A^*$, by the above proposition also $v \in A^*$. □

Corollary 2. *Let A be a code. If $uv \in A^*$ and $vu \in A^*$, $u \notin A^*$ and $v \notin A^*$, then, for any natural number n, $(uv)^n u \notin A^*$.*

Proof. Suppose that there exists a natural number n such that $(uv)^n u = u(vu)^n \in A^*$. Since $(uv)^n$ and $(vu)^n \in A^*$, it follows, by the above proposition, that also $u \in A^*$, in contradiction with the hypothesis of the corollary. □

Corollary 3. *Let A be a code and let $S = (f_1, f_2, ..., f_m)$ be any finite sequence of words of XX^* such that $f_k f_{k+1} \in A^*$ for any k $(1 \leqslant k < m)$. If there exist two indices i, j $(1 \leqslant i < j \leqslant m)$ such that $f_i \in A^*$ and $f_j \in A^*$, then, for all indices h such that $i \leqslant h \leqslant j$, $f_h \in A^*$.*

Proof. We first observe that, if, for a given index i $(1 \leqslant i \leqslant m)$, $f_i \in A^*$, then all the words of the form $f_{k'} f_{k'+1} \cdots f_{i-1} f_i f_{i+1} \cdots f_{k''-1} f_{k''}$ with $1 \leqslant k' \leqslant i \leqslant k'' \leqslant m$, belong to A^*. Thus, if $f_i, f_j \in A^*$, with $i < j$, we have that for any index h such that $i \leqslant h \leqslant j$,

$$f_i f_{i+1} \cdots f_{h-2} f_{h-1}, f_i f_{i+1} \cdots f_{h-1} f_h, f_{h+1} f_{h+2} \cdots f_{j-1} f_j, f_h f_{h+1} \cdots f_{j-1} f_j \in A^*.$$

It follows, by the above proposition, that $f_h \in A^*$. □

Verification of (a). Suppose that A is not very pure. Then, by Corollary 1, there exist two words $u, v \notin A^*$ such that $uv, vu \in A^*$. Let m be the natural number such that $uv \in A^m$. We prove that, for any natural number q, the pair $\{(uv)^q, (uv)^q\}$ of elements of A^{mq} is not synchronizing. Consider indeed the word $(vu)^{2q+1} \in A^*$. We have: $(vu)^{2q+1} = v(uv)^q(uv)^q u$ but, by Corollary 2, $v(uv)^q$ and $(uv)^q u \notin A^*$. Thus A has not finite synchronization delay.

Verification of (b). The proof of (b) and of Property 3 is obtained as a consequence of the Lemmas 1, 2 and 3 below. Let us first give some other definitions. If A is a code, for any word $g \in AA^*$, an A-factorization of g is a pair (g_1, g_2) of elements of AA^* such that $g_1 g_2 = g$. Being r and d natural numbers, denote by $V(r)$ and $U(d)$ the following conditions on a code A.

$V(r)$: For any $g \in A^r$ and for any $f, f' \in X^*$ such that $fgf' \in A^*$, there exists an A-factorization (g_1, g_2) of g such that $(fg_1, g_2 f')$ is an A-factorization of fgf'.

$U(d)$: For any sequence $S = (f_1, f_2, ..., f_i)$ of words of XX^*, $f_i f_{i+1} \in A^*$ for all i $(1 \leqslant i < d)$ implies $S \cap A^* \neq \emptyset$.

Lemma 1. *Let A be a code. If A satisfies $V(r)$, then A has synchronization delay $q < r$.*

Proof. The proof is by contradiction. If $q \geqslant r$, there exists a pair (a, a') of elements of A^{r-1} which is not synchronizing, that is there exist $f, f' \in X^*$ such that $faa'f' \in A^*$, but either $fa \notin A^*$, or $a'f' \notin A^*$, or both. Put $a = a_1 a_2 \ldots a_{r-1}$, $a' = a_r a_{r+1} \ldots a_{2r-2}$, $(a_i \in A)$, and $faa'f' = b_1 b_2 \ldots b_n$, $(b_j \in A)$. Consider the relation $fa_1 a_2 \ldots a_i =$

$b_1 b_2 \dots b_j$. If it is verified for $i < r-1$, it follows that $fa \in A^*$. Indeed $fa_1 a_2 \dots a_i \in A^*$ and $a_{i+1} a_{i+2} \dots a_{r-1} \in A^*$ imply $fa \in A^*$. If it is verified for $i > r-1$, it follows that $a'f' \in A^*$. Indeed $a_{i+1} a_{i+2} \dots a_{2r-2} f' = b_{j+1} b_{j+2} \dots b_n \in A^*$ and $a_r a_{r+1} \dots a_i \in A^*$ imply $a'f' \in A^*$. Since (a, a') is not a synchronizing pair, the above relation can be verified either for $i < r-1$, or for $i > r-1$, but cannot be verified in both cases. If the relation is verified only for $i < r-1$, then, considering the word $a_{r-1} a_r \dots a_{2r-2} \in A^r$, the condition $V(r)$ is not satisfied. If the relation is verified only for $i > r-1$, then, considering the word $a_1 a_2 \dots a_r \in A^r$, the condition $V(r)$ is not verified. □

Lemma 2. *Let A be a code that satisfies the conditions $F(p)$ and $U(d)$. Then A satisfies also the condition $V(pd)$.*

Proof. The proof is by contradiction. Assuming $F(p)$ verified, let us suppose that A does not satisfy $V(pd)$; there exist then a word $g = a_1 a_2 \dots a_{pd} \in A^{pd}$ (i.e. $a_i \in A$) and two words $f, f' \in X^*$ such that $fgf' = b_1 b_2 \dots b_n \in A^n$ (i.e. $b_j \in A$) and $fa_1 a_2 \dots a_i \neq b_1 b_2 \dots b_j$, for all i, j.

Let α be the mapping of the interval $[1, pd]$ in the interval $[1, n]$ defined as follows: for each $i \in [1, pd]$, $\alpha(i) = j \in [1, n]$ iff there exist two words b'_j, b''_j such that $b_j = b'_j b''_j$ and $fa_1 a_2 \dots a_i = b_1 b_2 \dots b_{j-1} b'_j$. α is not generally a one-to-one mapping. Consider the subset T of $[1, pd]$ such that $i \in T$ iff a_i is not an interior factor of $b_{\alpha(i)}$. By the condition $F(p)$, it follows that card $(T) = t \geqslant pd/p = d$. It can be also easily verified that the restriction of α to T is a one-to-one mapping of T onto $\alpha(T) \subset [1, n]$. Put

$$T = \{i_1, i_2, \dots, i_t\} \text{ with } i_1 < i_2 < \dots < i_t,$$

$$\alpha(T) = \{j_1, j_2, \dots, j_t\} \text{ with } j_1 < j_2 < \dots < j_t \text{ and } j_k = \alpha(i_k) \text{ for all } k\ (1 \leqslant k \leqslant t).$$

For each $k\ (1 \leqslant k \leqslant t)$, we have:

$$b_{j_k} b_{j_k+1} \dots b_{j_{k+1}-1} b_{j_{k+1}} = b'_{j_k} b''_{j_k} b_{j_k+1} \dots b_{j_{k+1}-1} b'_{j_{k+1}} b''_{j_{k+1}}$$
$$= b'_{j_k} a_{i_k+1} a_{i_k+2} \dots a_{i_{k+1}} b''_{j_{k+1}}.$$

Put

$$b_{j_k+1} b_{j_k+2} \dots b_{j_{k+1}-1} = \bar{b}_{j_{k+1}} \in A^*$$

and

$$a_{i_k+1} a_{i_k+2} \dots a_{i_{k+1}-1} = \bar{a}_{i_{k+1}} \in A^*.$$

Thus

$$\bar{b}_{j_{k+1}} b'_{j_{k+1}} b''_{j_{k+1}} \in A^*$$

and

$$b''_{j_k} \bar{b}_{j_{k+1}} b'_{j_{k+1}} = \bar{a}_{i_{k+1}} a_{i_{k+1}} \in A^*.$$

Consider now the sequence $S = (f_1, f_2, \dots, f_{2t})$ of words of XX^* defined as follows for each $k\ (1 \leqslant k \leqslant t)$:

$$f_{2k-1} = \bar{b}_{j_k} b'_{j_k},$$
$$f_{2k} = b''_{j_k}.$$

S satisfies the condition that for all m $(1 \leqslant m < 2t)$, $f_m f_{m+1} \in A^*$. According to the hypothesis that A^* is free, for each m, only one of the words f_m, f_{m+1} can belong to A^*. Indeed f_m and $f_{m+1} \in A^*$ for m odd, implies $\bar{b}_{j_k} b'_{j_k}$ and $b''_{j_k} \in A^*$ for some k, and the word $\bar{b}_{j_k} b'_{j_k} b''_{j_k}$ would have two different factorizations in words of A. f_m and $f_{m+1} \in A^*$ for m even, implies b''_{j_k} and $b_{j_{k+1}} b'_{j_{k+1}} \in A^*$ for some k, and the word $b''_{j_k} \bar{b}_{j_{k+1}} b'_{j_{k+1}} = \bar{a}_{i_{k+1}} a_{i_{k+1}}$ would have two different factorizations in words of A. Thus, by Corollary 3, it may exist at most one index m_0 $(1 \leqslant m_0 \leqslant 2t)$ such that $f_{m_0} \in A^*$. If $m_0 > t$, it follows that, considering the subsequence $S' = (f_1, f_2, \ldots, f_t)$ of S, the condition $U(t)$ is not satisfied. If $m_0 \leqslant t$, it follows that, considering the subsequence $S'' = (f_{t+1}, f_{t+2}, \ldots, f_{2t})$ of S, the condition $U(t)$ is not satisfied. Since $t \geqslant d$, if $U(t)$ if not satisfied clearly also $U(d)$ is not satisfied. The verification is concluded. □

Lemma 3. *Let A be a rational code and s the number of states of the minimal automaton recognizing A^*. If A is very pure, the condition $U(2s+2)$ is satisfied.*

Proof. The proof is by contradiction. If A does not satisfy $U(2s+2)$, there exists a sequence $S = (f_1, f_2, \ldots, f_{2s+2})$ of words of XX^* such that $f_k f_{k+1} \in A^*$ for all k $(1 \leqslant k < 2s+2)$, and $S \cap A^* = \emptyset$.

Since the right equivalence ρ_{A^*} induced by A^* has index s, we have certainly two elements f_i, f_j of S, with $i < j$ and $j - i$ even, such that $f_i \rho_{A^*} f_j$. By the definition of ρ_{A^*}, $f_i f_{i+1} \in A^*$ implies $f_j f_{i+1} \in A^*$. Since $j-i$ is even, i and j are both odd or both even; we have then:

$$f_{i+1} f_{i+2} \cdots f_{j-1} f_j \in A^*,$$
$$f_{i+2} f_{i+3} \cdots f_{j-2} f_{j-1} \in A^*.$$

By the last relation we also obtain:

$$f_{i+2} f_{i+3} \cdots f_{j-2} f_{j-1} f_j f_{i+1} \in A^*.$$

Put $u = f_{i+1}$ and $v = f_{i+2} f_{i+3} \cdots f_{j-1} f_j$. We have $uv, vu \in A^*$, but $u \notin A^*$. Hence A is not very pure. □

Proof of part (b) of the theorem is now obtained as follows. If A is a rational and very pure code and s is the number of states of the minimal automaton recognizing A^*, then by Lemma 3, $U(2s+2)$ is satisfied; if $F(p)$ is also verified, for some p, then, by Lemma 2, $V(2ps+2p)$ is verified and then, by Lemma 1, the synchronization delay q is less than $2p(s+1)$. Adding part (a) of the theorem to the above implications, Property 3 also follows. □

Acknowledgments

I wish to express my deep thanks to Jean-Francois Perrot for his criticism on the manuscript and many useful suggestions.

References

[1] A. H. Clifford and G. B. Preston, The Algebraic Theory of Semigroups, Vol. 2 (Am. Math. Soc., Providence, R.I., 1967).

[2] S. Eilenberg, Automata, Languages and Machines, Vol. A (Academic Press, New York1974).

[3] S. W. Golomb and B. Gordon, Codes with bounded synchronization delay, Information and Control 8 (1965) 355–372.

[4] S. W. Golomb, B. Gordon and R. L. Welch, Comma-free codes, Can. J. Math. 10 (1958) 202–210.

[5] A. Lentin and M. P. Schützenberger, A combinatorial problem in the theory of free monoids, in: Bose and Dowling, eds., Combinatorial Mathematics and its Applications (Univ. of North Carolina Press, Chapel Hill, N. Car., 1969) 128–144.

[6] R. McNaughton and S. Papert, Counter Free Automata (MIT Press, Cambridge, Mass.,1971).

[7] M. Nivat, Eléments de la théorie générale des codes, in: Caianiello, ed., Automata Theory (Academic Press, New York, 1966) 278–294.

[8] A. Restivo, Codes and aperiodic languages, in: Proceedings, GI. 1. Fachtagung über Automatentheorie und Formale Sprachen, Lecture Notes in Computer Science (Springer, Berlin, 1973).

[9] A. Restivo, On a question of McNaughton and Papert, Information and Control 25 (1974) 93–101.

[10] M. P. Schützenberger, Une théorie algébrique du codage, Séminaire Dubreil–Pisot, 1955–1956, Paris.

[11] M. P. Schützenberger, Sur une propriété combinatoire des algèbres de Lie libres pouvant être utilisée dans un problème de mathématique appliquée, Séminaire Dubreil–Pisot, 1958–1959, Paris.

[12] M. P. Schützenberger, Codes à longueur variable, Cours a l'école d'été de l'OTAN sur les méthodes combinatoires en théorie du codage, Royan (1965).

[13] M. P. Schützenberger, On a factorization of free monoids, Proc. AMS 16 (1965) 21–24.

[14] M. P. Schützenberger, Sur certaines opérations de fermeture dans les langages rationnels, Communication au Convegno di Informatica Teorica, Roma, Symp. Math. 18 (1974).

[15] G. Viennot, Factorisations régulières des monoïdes libres et algèbres de Lie libres, C. R. Acad. Sci. Paris 277 (A) (1973) 493–496.

[16] G. Viennot, Algèbres de Lie libres et monoïdes libres, Thèse de Doctorat d'Etat, Université Paris VII (1974).